Direct determination of molecular constants from rovibronic spectra with genetic algorithms

J. A. Hageman and R. Wehrens
Laboratory of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

R. de Gelder
Department of Inorganic Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

W. Leo Meerts
Department of Molecular and Laser Physics, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

L. M. C. Buydens[a)]
Laboratory of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

(Received 3 February 2000; accepted 11 August 2000)

It is shown that a new procedure, based on genetic algorithms (GA's), can be used for direct determination of molecular constants, in particular rotational constants, from rovibronic spectra. This new approach only requires an estimate of the acceptable range of the parameters. The power of the method is demonstrated on the rotationally resolved fluorescence spectra of indole, indazole, benzimidazole, and 4-aminobenzonitril. A rigid asymmetric rotor Hamiltonian is used to calculate the theoretical spectra. The GA matches the generated spectra with an experimental spectrum with the use of a new method for spectra comparison. This spectra comparison function is able to deal with frequency shifts which are caused by (small) changes in the rotational constants and it yields better results in comparison with traditional spectra comparison methods, like RMS. In addition, the robustness of the method is tested. © *2000 American Institute of Physics.* [S0021-9606(00)00342-1]

I. INTRODUCTION

Rotational constants are an important tool in determining the spatial geometry of molecules. These constants give access to intramolecular and intermolecular bond lengths and their changes upon excitation. Rotational constants can be obtained from a large variety of methods, among others microwave spectroscopy, IR Fourier transform, diode laser spectroscopy and high resolution laser induced fluorescence (LIF) excitation spectra. Especially the last three methods deal with a two state problem, either two vibrational or two electronic states. The resolution of such spectra is such that individual rotational transitions can be observed and the spectra contain large number of lines. Usually, to determine molecular rotational constants a spectrum is simulated using a model (for instance an asymmetric rotor Hamiltonian) which uses rotational constants estimated from other experiments or from *ab initio* calculations and the appropriate selection rules. The spectrum is calculated and compared with the experimental one. In what we will call the classical method, an initial assignment in terms of theoretical quantum numbers of transitions is made. In a least-squares-fit procedure the molecular parameters are adjusted to reproduce the assigned lines. The assignments are refined and the process is repeated until all lines in the spectrum are successfully reproduced.[1] The quality of the fit is, amongst other factors, dependent on the sophistication of the used model. The determination can be facilitated and speeded up by using reasonably accurate estimates of the molecular constants although this is not crucial.

Recently, attempts to automate the interpretation of rovibronic spectra have been undertaken. Automation becomes increasingly important when spectra become more difficult to interpret and/or prior knowledge about the molecule is little or lacking. The group of Neusser *et al.*[2] used a procedure which directly fits the experimental data, without any preceding assignment of lines, with the help of the so-called correlation automated rotational fitting algorithm. This algorithm still relies on accurate initial estimates of the rotational constants obtained from other experiments. Their experimental data were also analyzed by assigning lines and using a least-squares-fit procedure. They concluded, by visual inspection, that the correlation method yielded better results.[2] Unfortunately, the method still has limited applicability.

The approach of fine-tuning the parameters of the Hamiltonian model so that the theoretical spectrum is in close agreement with the experimental one, can be seen as an optimization problem. The process of determining molecular constants can be automated with global optimization methods like simulated annealing (SA),[3] Tabu search (TS)[4] or

a)Author to whom correspondence should be addressed.

genetic algorithms (GA's). In this paper it is shown that a GA with a specially developed fitness function is very successful in directly determining the molecular constants from LIF spectra. This is done without using any initial estimates of these constants, except their global limits. This new approach is demonstrated for four rotationally resolved (LIF) spectra from indole, indazole, imidazole, and 4-aminobenzonitril (4-ABN). The spectra were measured by Berden *et al.*[5,6] The essence of the analysis of Berden *et al.* was an assignment of quantum numbers of the initial and final states of the transitions in the spectrum. In a second step an overall fit of the intensities was carried out in which only the intensity parameters were determined. By carefully adapting the parameters Berden *et al.* succeeded in minimizing the difference between the experimental and simulated spectrum and obtained the complete set of molecular constants.

In the next section, a description is given of the parameters that appear in the Hamiltonian model, the use of GA's and the new method for comparing spectra. It will be shown in Sec. III that a GA is very capable of determining the molecular parameters that reproduce the experimental spectra. In addition, the robustness of this GA-based method will be assessed by artificially deteriorating the quality of the data. It is shown that the method is quite robust and, therefore, widely applicable.

II. THEORY

A. Model representation

Given a set of molecular parameters, a theoretical rovibronic spectrum can be calculated using a rigid asymmetric rotor Hamiltonian. All experimental spectra analyzed in this paper are fitted to this type of calculated spectra. It is assumed that, if a theoretical spectrum matches the experimental one, the parameters used to calculate the spectrum are very close to the true values. Since a discussion of the theory of the rigid asymmetric rotor Hamiltonian is beyond the scope of this paper, we will suffice to say that all rotational levels of the molecules under study can be calculated with this model.[5,6] The important details of the model are described briefly below. It contains 13 parameters, which are optimized by the GA. They can be divided in five groups.

(1) Six rotational constants. Three parameters (A'',B'',C'') describing the ground state and three parameters $(\Delta A, \Delta B,$ and $\Delta C)$ describing the difference between the ground and excited state values, $\Delta A=(A'-A'')$, etc. Here the double and single primes label the ground and excited states, respectively. These parameters are responsible for the positions of the transition frequencies.
(2) A frequency shift parameter (ν). This parameter shifts the whole spectrum relative to an arbitrary zero point.
(3) Three parameters that describe the relative intensities of the transitions (T_1,T_2,W). A three-parameter two-temperature model has been used.[5] By definition, T_2 must be higher that T_1. W is a weighting factor.
(4) The direction of the electronic transition moment of the electronic excitation (θ) and a parameter (θ_T), which is the angle between the principal axes systems in the ground and in the excited states. θ_T is not optimized in this approach as it influences only a very small number of lines (<10) and can only be determined by visual inspection of the appropriate lines. See, for an example, Fig. 5 in Ref. 5.
(5) The linewidth $(\Delta\nu)$ of lines in the spectrum. In Ref. 5 it is shown that the transitions in all four spectra are best described by a Lorentzian profile. However, this is not an essential limitation for the present discussion.

B. Genetic algorithms

GA's are a special class of global optimizers, based on the theory of evolution. A GA is able to minimize (or maximize) a function $G(x)$, where x represents a parameter vector, by searching the parameter space of x for the optimal solution. Several general steps can be distinguished that are identical for all GA's.

(1) Initialization: GA's do not operate on an individual solution for searching the parameter space but rather on a group of solutions (called population) at a time. A solution, called string or chromosome, consists of several parts, called genes. Each part is a parameter which needs optimization. All initial solutions are set to random values. In the present examples each chromosome contains 12 genes which are the 12 parameters of the rigid Hamiltonian model.
(2) Evaluation: All strings in the population are evaluated by an objective function. This results in a measure of quality of the string, expressed in a single number. The evaluation function is usually tailor made for the specific GA application. It should be able to discriminate between good and bad strings in a given generation, to allow selection in a later phase.
(3) Stop: A stop criterion is checked.
(4) Selection: A percentage of the best strings in a population is selected and placed in the next generation.
(5) Recombination: To form the new population, new solutions are created by combining two existing solutions (parents) to yield two different ones (children). This is called crossover.
(6) Mutation: Genes on a string in the new population are selected randomly and modified. An example of a mutation operator is addition of a (small) random number. To prevent the search from random behavior the probability of mutation is usually chosen to be quite low.

Several parameters, for instance the rate of crossover and mutation, regulate the performance of the GA and each specific problem has its own specific set of parameters for which the GA performs at its optimum. This so-called meta-optimization of the GA parameters can be tedious and can be considered a disadvantage of GA in general.[7] In this paper it is shown that one set of GA parameters can successfully be used for estimating molecular constants of indole, indazole, benzimidazole, and 4-ABN, so it is not necessary to repeat this meta-optimizing for each new compound. The most important advantage of the GA approach is that it does not need prior knowledge of the molecular constants. All that is required, is an estimate of the accessible range for each param-

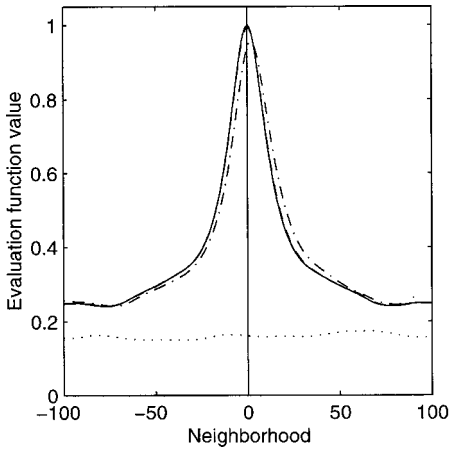

FIG. 1. Correlogram of the calculated spectrum of indole. Autocorrelogram: solid line, Cross correlogram: dashed line (A'' increased by 1.0 MHz), dashed–dotted line (ΔA increased by 1.0 MHz), and dotted line (calculated spectrum of benzimidazole).

eter. The narrower this range is chosen the faster the optimization will be. These ranges can be chosen, for instance, on physical grounds or be adapted from similar molecules known from the literature.

Some literature is available about GA's. For an introduction to the subject Ref. 8 or for a more sophisticated level Ref. 9 are very well suited.

C. Evaluation or objective function

The parameters on each string are used in the rigid asymmetric rotor Hamiltonian model to calculate a theoretical spectrum. The similarity between the calculated spectrum and experimental spectrum has to be expressed in a single number. Several methods are known to compare spectra. The more traditional methods perform a pointwise comparison between two spectra and express this as the sum of the squared differences.[9,10] More sophisticated comparison methods include a comparison of the neighborhood to deal with peak shifts.[11]

Our initial attempts clearly demonstrated the inability of evaluation functions based on a sum of squared differences to differentiate reliable between spectra originating from nearly identical sets of parameters. Other approaches, based on peak picking and minimizing the distance to neighboring peaks in both spectra, failed as well. Moreover, since the relative position of peaks can change dramatically, one is never sure if the correct peak pairs are compared. With these types of evaluation functions, similar spectra with shifts in peak positions will not properly be recognized as similar. An improvement over the RMS-based evaluation function is the correlation coefficient *Cr*:

$$Cr=\frac{\Sigma_{x=0}^{x=k}(f(x)-\langle f\rangle)(g(x)-\langle g\rangle)}{\sqrt{\Sigma_{x=0}^{x=k}(f(x)-\langle f\rangle)^2\Sigma_{x=0}^{x=k}(g(x)-\langle g\rangle)^2}}. \tag{1}$$

Here $\langle f\rangle$ and $\langle g\rangle$ are the average intensity value of spectrum f and g, respectively, $f(x)$ and $g(x)$ are the spectra f and g with length k. Further improvement is obtained if the average value is no longer subtracted from each point and (a possible) elevated baseline is removed. This ensures that all signals present in the spectrum are originating from peaks. This yields an evaluation function $\cos\gamma$:

$$\cos\gamma=\frac{f\cdot g}{\|f\|\|g\|}. \tag{2}$$

Here $f\cdot g$ is the dot product of the experimental (f) and calculated (g) spectrum, $\|f\|$ and $\|g\|$ represent the length of spectrum f and g, respectively. $\cos\gamma$ ranges from $\langle -1|1\rangle$. For equal spectra $\cos\gamma=1$.

Equation (2) does not take into account small frequency shifts in the peak position. The evaluation function of Eq. (2) can be improved to deal with shifts if a cross correlation function is used:

$$C_{fg}(r)=\frac{\Sigma_{x=0}^{x=k}f(x)\cdot g(x+r)}{\|f\|\cdot\|g\|}. \tag{3}$$

The cross correlation function compares two spectra shifted by r. In order to deal with end points the sum should run from $-\infty$ to $+\infty$. Formally this can be realized by adding to the spectra points of zero intensity. In this way, the normalization by $\|f\|$ and $\|g\|$ is properly defined. Figure 1 shows several C_{fg} with r ranging from $[-100,100]$. The solid line is the autocorrelogram where both f and g are the original calculated spectrum of indole from Ref. 5. The dashed and the dashed–dotted line are cross correlograms of the calculated spectra of indole with two calculated spectra in which A'' and ΔA, respectively, are slightly changed by 1.0 MHz. The dotted line is the cross correlogram of the calculated spectrum of indole with the calculated spectrum of benzimidazole from Ref. 5. It can be seen from Fig. 1 that no shift whatsoever of the benzimidazole spectrum will significantly increase C_{fg}, indicating no correlation at all between the

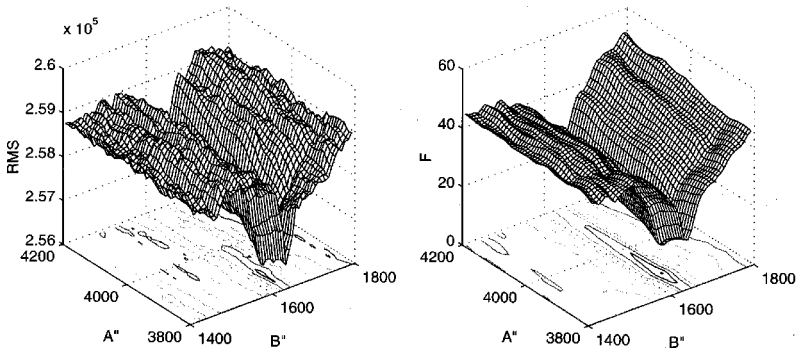

FIG. 2. Difference in error landscape between a RMS-based evaluation function (left) and one based on Eq. (6) (right).

spectra of indole and benzimidazole. Although the solid, dashed–dotted, and dashed lines originate from spectra calculated with nearly identical parameters, the change in $C_{fg}(0)$ [which equals Eq. (2)] is quite large. This implies that almost identical spectra may have quite different values for $C_{fg}(0)$. However, the area under the curve can be used as a convenient measure if a suitable weight function is used.

To penalize larger shifts, Eq. (3) is modified by introducing a weight function $w(r)$:

$$w(r)=1-\frac{|r|}{l}. \tag{4}$$

The parameter l defines the width of the neighborhood that is taken into account, typically in the order of 100 data points in the current work. Several weight functions were tested, including the sigmoidal function from Ref. 11. Eventually the simple triangle function [Eq. (4)] is used, because it depends on only one parameter. The sigmoidal function showed no improvement over Eq. (4).

The final overlap function is obtained by integrating Eq. (3) multiplied by the weight function and normalizing between 0 and 1:

$$C_{fg}^{ws}=\frac{\Sigma_{r=-l}^{r=l}C_{fg}(r)w(r)}{\sqrt{\Sigma_{r=-l}^{r=l}C_{ff}(r)*w(r)}*\sqrt{\Sigma_{r=-l}^{r=l}C_{gg}(r)*w(r)}}. \tag{5}$$

For two identical spectra C_{fg}^{ws} is 1 and for two distinctly different spectra C_{fg}^{ws} is close to zero. The final evaluation function used in the GA calculations is defined as

$$F=100*(1-C_{fg}^{ws}) \tag{6}$$

and its value is minimized.

Error landscapes of an RMS-based evaluation function and F are plotted in Fig. 2. In both plots A'' and B'' are varied over a grid covering the complete range, while the remaining parameters are held fixed. The effect of Eq. (6) clearly shows a more smooth error landscape, which reduces the number of local minima.

A more detailed discussion and comparisons with other methods for the assessment of similarity between one-dimensional spectra can be found in the work of De Gelder *et al.*[12]

III. EXPERIMENT

The spectra of indole, indazole, benzimidazole, and 4-ABN are shown in Fig. 3. The spectra of indole and benzimidazole contain 65 536 equidistant data points, the spectrum of indazole 61 821 data points and the spectrum of 4-ABN contains 40 972 data points. All 12 parameters were coded as 10-bit gray binary numbers. T_2 is coded on the string as α, with $T_2=\alpha*T_1$ and $\alpha>1$. The calculated spectra always contain the same number of data points as the corresponding experimental ones. The optimal settings of the GA were determined by trial and error and based on previous experience using the experimental spectrum of benzimidazole and are shown in Table I.

The optimal size of the neighborhood in Eq. (5) has been established from several experiments. The optimal value for

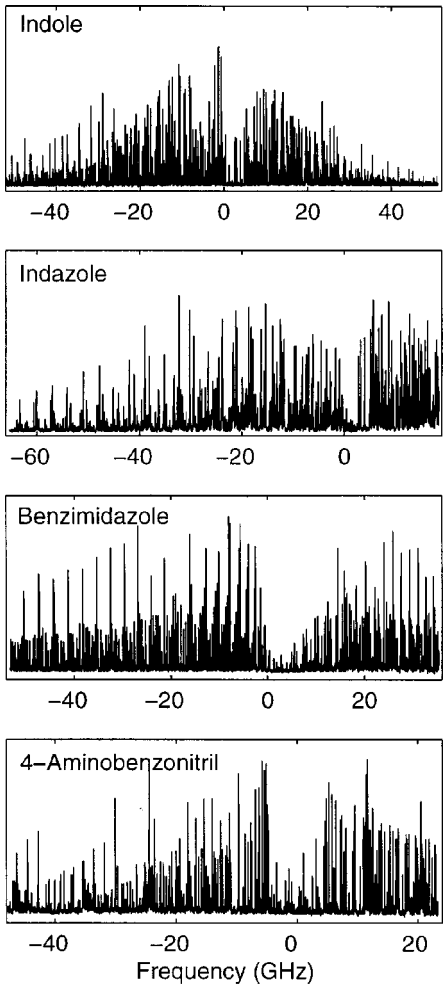

FIG. 3. High resolution LIF spectra of indole, indazole, benzimidazole and 4-ABN. In all cases the absolute frequency is set to 0.0 according to Refs. 5 and 6. The intensity on the vertical scale is in arbitrary units.

l was 100 data points. A larger range also results in a correct solution but leads to longer run times. For a significantly smaller range no correct solution is obtained indicating that the inclusion of neighborhood information is crucial. After establishing the optimal settings, the experimental spectra of indole, indazole, benzimidazole and 4-ABN were fitted using boundary constraints as given in Table II. The duration of a run has been set to 500 generations, long enough to converge

TABLE I. GA settings.

Setting	Value
Maximum number of generations	500
Population size	300
Elitism	150
Crossover type	Two-point crossover
Crossover probability	0.85
Mutation type	New random value within boundaries
Mutation probability	0.05
Selection type	Probabilistic
Fitness type	Raw[a]

[a]Fitness value increases inversely proportional with evaluation value of a string.

TABLE II. Boundary constraints for all 12 parameters used for Indole, Indazole, and Benzimidazole and 4-ABN.[a]

Parameter	Boundary constraints indole and benzimidazole	4-ABN
A''	3800–4200	5000–6000
B''	1400–1800	800–1200
C''	800–1400	600–1000
T_1	1–6[b]	1–6
T_2[d]	1.5–5	1.5–5
W	0–1	0–1
θ	0°–90°	90°, fixed[e]
ν	−300–300[c]	−5000–5000
ΔA	−200–0	−400–400
ΔB	−50–0	−100–100
ΔC	−50–0	−100–100
$\Delta\nu$	10–40	10–90

[a]Rotational constants in the ground state are indicated by A'', B'', and C''. Rotational constants in the excited state are given by their deviations from the ground state (ΔA, ΔB, and ΔC). $\Delta\nu$ is linewidth of the Lorentzian peaks. Rotational constants, ν and $\Delta\nu$ are in MHz, T_1, and T_2 in K.
[b]Range is 2–8 for the spectrum taken from benzimidazole.
[c]The frequency of the origin (ν) is set to zero. The area of deviation is taken to be ±10% of the reported value from Refs. 5 and 6.
[d]$T_2=\alpha*T_1$ where α has been optimized with the constrained $\alpha>1$.
[e]Determined by the geometry of the molecule.

to a minimum. All runs were repeated five times with different random generator seeds to exclude lucky and/or unlucky runs.

The robustness of the GA method was investigated in a number of runs. We investigated the influence of (high) noise levels, increased linewidths, and the total number of points in a spectrum. Synthesized spectra of indole and benzimidazole were modified with different levels of normally distributed (white) noise, increased linewidths, and a combination of these two factors. Spectra with a reduced number of data points were also investigated. Figure 4 shows parts of the spectrum of indole with (a) a signal-to-noise level (S/N) of 10 (for the peak with the largest intensity), (b) a linewidth of 90 MHz, and (c) a S/N of 10 combined with a linewidth of 90 MHz. The spectrum with a combination of large linewidths and low S/N can be considered as very extreme.

All GA calculations were performed with the GA library PGAPack version 1.0,[13] which can run on parallel processors. PGAPack and the evaluation function are written in ANSI-C, the rigid asymmetric rotor Hamiltonian function was written in Fortran. All calculations were performed on a Sun–Ultra-Enterprise–10000 with 24 processors each running at 333 MHz. With 16 processors, the average runtime was about half an hour for 500 generations and 65 536 data points. In practice this run time can be reduced drastically, because often runs converge to their final solution long before the maximum number of generations is reached. If the boundaries are taken narrower, run times can be further reduced because runs will converge even earlier. This will also lead to increased reproducibility and this decreases the need for more replicated runs. On a single processor (500 MHz) one complete analysis can be executed in about 12 hours.

IV. RESULTS AND DISCUSSION

Table III shows the 12 parameters for all four experimental spectra as they resulted from the GA, together with the results of a re-evaluation of the spectra reported in Ref. 5 (indole, indazole, and benzimidazole) and Ref. 6 (4-ABN) using the classical methods. The molecular constants from Ref. 5 are averages from multiple spectra and were determined using very accurate ground rotational constants from microwave experiments. Results reported in Table III are based on a spectral analysis of the same spectrum as used for the GA method and the ground rotational constants were also determined from that spectrum. The values obtained in the present GA approach are inclose agreement with those from the classical method. For most of the parameters the results are within the experimental error. If the results are outside the error, the deviations are very small. These deviations are

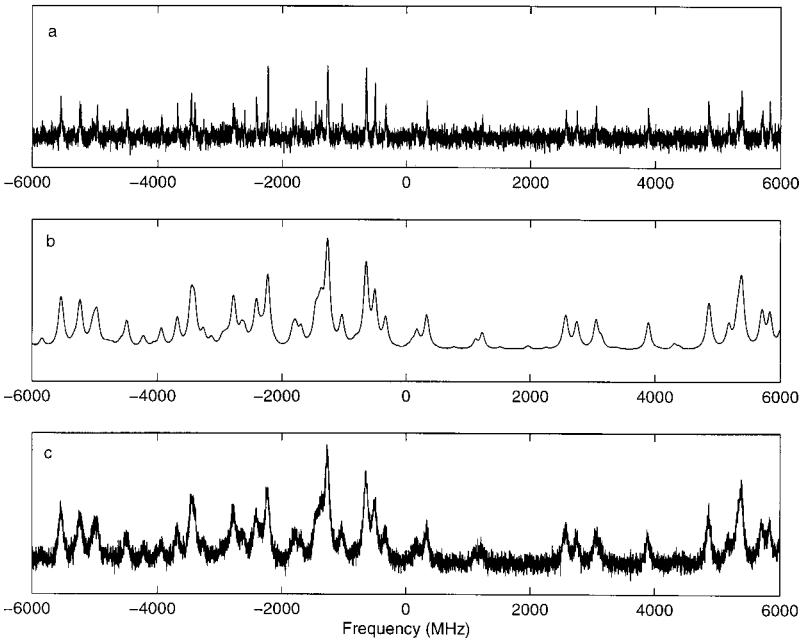

FIG. 4. Synthesized spectra of indole with (a) S/N = 15 (for strongest line), (b) $\Delta\nu=90$ MHz, and (c) S/N = 15 (for strongest line) together with a $\Delta\nu=90$ MHz. The intensity on vertical scale is in arbitrary units.

TABLE III. Results from GA runs for indole, indazole, benzimidazole, and 4-ABN.[a]

	Indole		Indazole	
	GA	Ref. 5[b]	GA	Ref. 5[b]
A''	3879.8	3880.7 (1.0)	3979.9	3979.2 (0.8)
B''	1637.0	1637.5 (0.4)	1633.8	1633.9 (0.3)
C''	1151.3	1152.1 (0.4)	1158.4	1158.6 (0.3)
T_1	2.22	1.50	2.60	2.60
T_2	7.93	5.03	7.35	8.18
W	0.1	0.22	0.23	0.19
θ	37.4°	±38.3	62.3°	62.2°
ν[c]	0.78	0.0 (1.6)	−1.7	0.0 (1.7)
ΔA	−134.70	−134.66 (0.09)	−102.44	−102.30 (0.09)
ΔB	−18.08	−17.96 (0.18)	−29.23	−29.20 (0.13)
ΔC	−20.72	−20.77 (0.32)	−23.31	−23.20 (0.28)
$\Delta\nu$	16.158	20.05	26.452	32.75
	Evaluation values			
Best	4.18		0.68	
mean	4.24		0.74	
std. dev.	0.08		0.06	
	Benzimidazole		4-ABN	
	GA	Ref. 5[b]	Ga	Ref. 6
A''	3929.0	3930.5 (1.0)	5579.7	5579.3 (0.5)
B''	1679.2	1679.5 (0.2)	990.23	990.26 (0.09)
C''	1177.1	1176.7 (0.2)	841.45	841.39 (0.08)
T_1	5.63	4.88	2.63	3
T_2	21.52	20.0	4.56	
W	0.42	0.42	0.84	
θ	22.1°	±22.0°	0°	0°
ν[c]	1.04	0.0 (1.64)	−1.61	0.0
ΔA	−155.62	−155.70 (0.03)	−315.54	−316.61 (0.06)
ΔB	−15.30	−15.37 (0.08)	10.66	10.849 (0.003)
ΔC	−21.41	−21.31 (0.13)	0.29	0.095 (0.001)
$\Delta\nu$	19.33	19.45	16.16	26
	Evaluation values			
Best	0.65		1.2	
mean	0.71		14.7	
std. dev.	0.06		13.9	

[a]Values from Refs. 5 and 6 are listed in the respective columns. Rotational constants, ν and $\Delta\nu$ are in MHz, T_1 and T_2 in K.
[b]Results in this column differ partly from those reported in Ref. 5. See the text for details.
[c]The absolute frequency of the origin is given as the deviation from the reported value from Refs. 5 and 6.

probably caused by the lack of precision of a GA. It is known that GA's can locate the global minimum but that they are not as precise as, for instance, local optimizers.

The GA method gives no information about the accuracy of the best fit parameters. However, it should be possible to assign quantum numbers to the experimental transitions after a GA fit. The experimental errors can then be estimated by performing a classical calculation like in Ref. 5, where it is no longer necessary to go through the sometimes tedious assignment process.

All GA runs were repeated five times with different seeds for the random number generator and the solution with the lowest evaluation values are shown in Table III. Results from Ref. 5 can be expected to be more accurate because the ground rotational constants were determined by microwave experiments which are more precise.

The parameters that describe the relative intensity of a transition (T_1, T_2, W) have different values in comparison with the reported values from Ref. 5. (Reference 6 used a one-temperature model so this cannot be compared with the present results.) The deviation is due to the fact that for these parameters several sets can be used with equal spectral intensities as a result.

The GA was able to find the correct solution for the indole, indazole, and benzimidazole spectra in all five replicated runs. For the 4-ABN data, the correct solution was only found in two of the five cases, as shown in Fig. 5. The cause of the reduced reproducibility of the 4-ABN run is probably due to larger boundary constraints, which makes it more difficult for the GA to locate the correct solution.

The absolute evaluation function values do not reach the same level for the four compounds. This is due to noise level, linewidth, and total number of data points in a particular spectrum. High noise levels intrinsically give rise to large evaluation function values. However, the minimum reached in each case is the global minimum irrespective of the absolute evaluation value. The similarity between all four experimental and the corresponding calculated spectra is very high. As an example this is shown in Fig. 6 for a representative area of the spectrum of indole.

V. APPLICABILITY OF THE GA METHOD TO PARTLY RESOLVED SPECTRA

Figure 7 shows results for synthesized spectra of indole and benzimidazole with increased noise levels, linewidths, and spectral resolutions. Again, the GA runs were repeated five times with different seeds for the random number generator. The best set of parameters found in these runs was used to generate spectra which are free of noise, have normal linewidths, and the same number of data points as the source spectra. The source spectra and the generated spectra are then compared with the evaluation function F. The evaluation values calculated in this way can directly be used to compare the quality of the different fits since the effects of added noise and linewidths is removed from the evaluation function. Figure 7 demonstrates the results for indole and benzimidazole.

In both cases, all modifications to the calculated spectra lead to an increase in evaluation value and thus in a deterioration in quality of the solution. However, the effect of the increased linewidths is somewhat less and more constant than the effects of other contributions. If the noise levels increase, the quality of the solutions decreases. The combination of both increased linewidths and high noise levels does not lead to further deterioration of the best solution. A decrease of the number of data points (where the frequency is kept constant) only shows an effect on the benzimidazole spectrum. For a smaller number of data points, the solutions become worse. This is due to the fact that spectral information gets lost if the distance between two successive data points becomes too large. Although the quality of the obtained parameters deteriorated, the rotational constants are hardly influenced by the elevated noise levels. The deviations are mostly found in T_1 and T_2 and in θ. Because one is mostly interested in the rotational constants the method can be considered quite robust for the determination of these parameters.

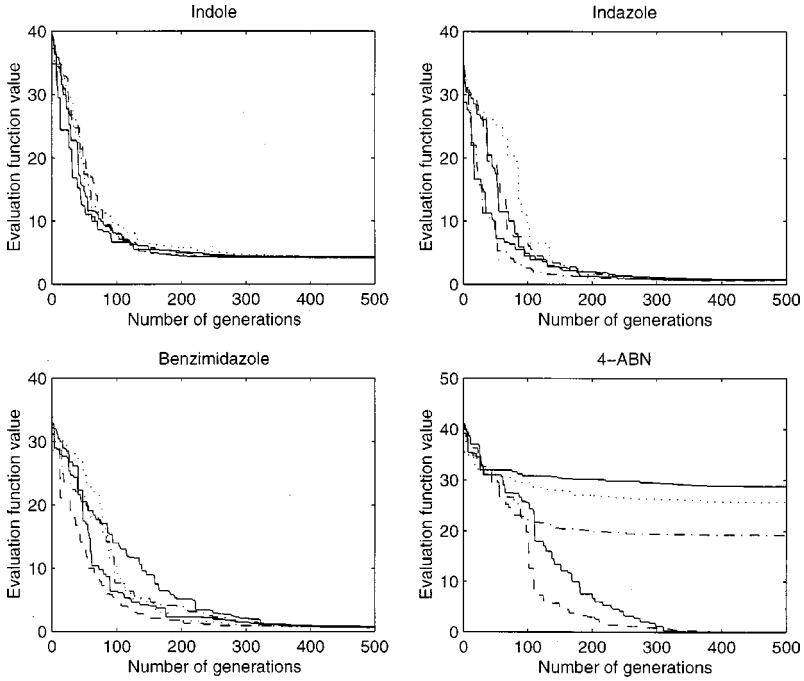

FIG. 5. Progression of the best solution during a run for indole, indazole, benzimidazole, and 4-ABN.

VI. CONCLUSION

The automated interpretation of high resolution spectra becomes of great importance if the interpretation by other methods is no longer feasible, too time consuming or more a routine matter. In the approach presented in this paper, only knowledge of the range of the parameters is needed for the deduction of molecular constants. In general, feasible ranges can be given and may even be quite large. The meta optimizing can be tedious for GA's. However, in the present case it is demonstrated that one set of GA settings suffices to retrieve the molecular constants from different rotationally resolved spectra. The success of the GA method crucially depends on the newly developed evaluation function. Other, more standard, evaluation functions lead to no results.

The problem of spectrum comparison in this particular application is related to peak shifts which are caused by small changes in the rotational constants. This makes it necessary to include a comparison of the neighborhood of a given point in the spectrum. All attempts based on the sum of squared differences without considering the neighborhood of points failed, precisely because these criteria do not properly deal with peak shifts. This demonstrates that a special

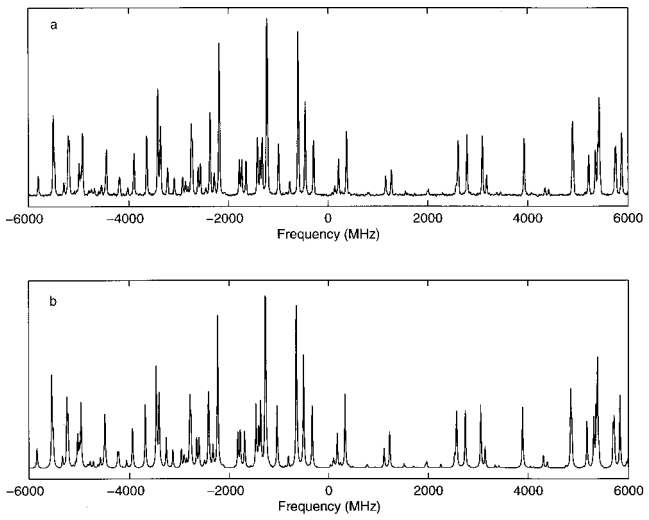

FIG. 6. Representative area of the experimental (a) and calculated (b) spectrum of indole. The intensity on the vertical scale is in arbitrary units.

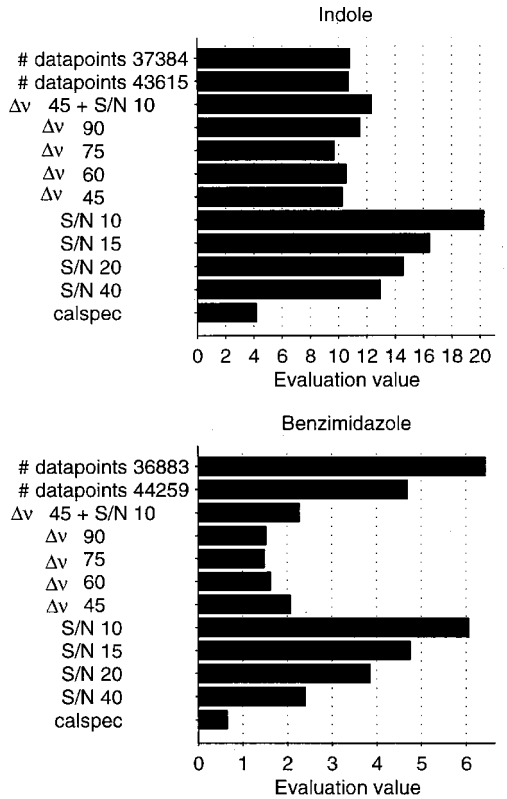

FIG. 7. Influence of noise (S/N), linewidth ($\Delta\nu$) and the total number of data points in a spectrum on the best solution found of indole (top) and benzimidazole (bottom). Calspec indicates the spectrum which fits with the experimental one best.

tailor-made evaluation function is crucial to obtain any results. It shows that, apart from an optimization of the settings of the GA, GA's in combination with a standard evaluation function cannot be used as a black box to solve any optimization problem.

The GA method is quite robust. It is insensitive to large linewidths in the spectrum, and only at very high noise levels do the results deteriorate. It is shown that the GA is able to use all information present in the spectrum and therefore its performance increases with the number of data points. The method of matching experimental data (represented as a vector) with simulated model data by optimizing model parameters with a GA can be successfully used in other fields, especially with the newly developed evaluation function.

ACKNOWLEDGMENTS

The authors thank Dr. Giel Berden for a critical reading of this paper.

[1] G. Berden *et al.*, J. Chem. Phys. **104**, 3935 (1996).
[2] R. Helm, H.-P. Vogel, and H. Neusser, Chem. Phys. Lett. **270**, 185 (1997).
[3] *Adaption of Simulated Annealing to Chemical Optimization Problems*, Vol. 15 of *Data Handling in Science and Technology*, edited by J. Kalivas (Elsevier, Amsterdam, 1995).
[4] F. Glover and M. Laguna, *Tabu Search* (Kluwer Academic, Dordrecht, 1998).
[5] G. Berden, W. Meerts, and E. Jalviste, J. Chem. Phys. **103**, 9596 (1995).
[6] G. Berden, J. V. Rooy, W. Meerts, and Z. Zachariasse, Chem. Phys. Lett. **278**, 373 (1997).
[7] R. Wehrens, E. Pretsch, and L. Buydens, Anal. Chim. Acta **388**, 265 (1999).
[8] R. Wehrens and L. Buydens, TrAC, Trends Anal. Chem. **17**, 193 (1998).
[9] K. Harris, R. Johnston, and B. Kariuki, Acta Crystallogr., Sect. A: Found. Crystallogr. **54**, 632 (1998).
[10] J. Dods, D. Gruner, and P. Brumer, Chem. Phys. Lett. **261**, 612 (1996).
[11] H. Karfunkel *et al.*, J. Comput. Chem. **14**, 1125 (1993).
[12] R. de Gelder, R. Wehrens, and J. A. Hageman, J. Comp. Chem. (accepted).
[13] D. Levine, PGAPack V1.0, PGAPack can be obtained from anonymous ftp from ftp://ftp.mcs.anl.gov/pub/pgapack/pgapack.tar.Z.